

Non-Smooth Variational Data Assimilation with Sparse Priors

A.M. Ebtehaj^{1,2}, E. Foufoula-Georgiou¹, S.Q. Zhang³ and A.Y. Hou³

¹Department of Civil Engineering, Saint Anthony Falls Laboratory, University of Minnesota

²School of Mathematics, University of Minnesota

³NASA Goddard Space Flight Center, Greenbelt, Maryland

This paper proposes an extension to the classical 3D variational data assimilation approach by explicitly incorporating as a prior information, the transform-domain sparsity observed in a large class of geophysical signals. In particular, the proposed framework extends the maximum likelihood estimation of the analysis state to the maximum a posteriori estimator, from a Bayesian perspective. The promise of the methodology is demonstrated via application to a 1D synthetic example.

1 Introduction

Data Assimilation has played a central role in improving the forecasting ability of hydro-meteorologic, climatic and oceanic modeling systems. The basic idea is to consistently fuse the observations of the prominent state variables (e.g., wind, temperature, pressure) or physical states (e.g., cloud moisture, precipitation), iteratively in time, into the knowledge of a Numerical Prediction Model (NPM) to reduce the estimation uncertainty of the state variables of interest. Data assimilation methods typically use the observations to update the current a priori model estimates of the states (*background*) and produce an a posteriori state (*analysis*) to be used for prediction of the next time step (*forecast*).

Data assimilation primarily stems from the least-squares estimation in the statistical sense. Basically, the available methods can be divided into two major categories. The first, is the group of recursive (least-squares) *filtering* methods which essentially exploit the temporal evolution of some statistical characteristics of the system (e.g., covariance) to efficiently track the optimal states sequentially in time (e.g., Kalman filter driven data assimilation methods). The second category, the so-called *variational* methods, relies on a batch

mode direct optimization at each instant of time when the observations become available (e.g., 3D or 4D variational approaches). We remark here that, these two apparently distinct approaches often share similar mathematical concepts and are quite equivalent in many cases; however, with different implementation strategies in practice. In this paper, we restrict our attention to the second group of assimilation methods and in particular the more primitive 3D variational (3D-VAR) formulation. For a thorough review of the historical evolution of the data assimilation techniques the reader is referred to [Talagrand and Courtier \(1987\)](#), [Ghil and Malanotte-Rizzoli \(1991\)](#), [Daley \(1993\)](#), [Bouttier and Courtier \(2002\)](#), [Kalnay \(2003\)](#), [Zhou et al. \(2006\)](#), [Evensen \(2007\)](#), and references therein.

The classical variational data assimilation typically involves solving a *smooth* optimization problem in which the solution has a minimum weighted Euclidean distance to both observation and background estimates where the weights are dictated by the pair of model and observation error covariance matrices. From a statistical estimation point of view this procedure is equivalent to the Maximum Likelihood (ML) estimation of the unknown state in a Gaussian noise (error) environment. In this classical formulation no *a priori* assumption is explicitly taken into account about the underlying structure of the analysis state.

Natural signals can typically be projected onto transform domains (e.g., Fourier, Discrete Cosine, Wavelet) in which a large fraction of the representation coefficients is very close to zero and only a few of them are significant, a signature typically referred to as “sparsity”. For instance, the wavelet transform of piece-wise smooth natural signals with occasional rapid variations often translates to non-Gaussian heavy tail distribution of the wavelet coefficients with a concentrated probability mass around zero.

Here, we propose a new formalism for variational data assimilation which explicitly incorporates the underlying sparsity in the analysis state as an *a priori* knowledge. In a very simple example we demonstrate how this *a priori* knowledge can stabilize and make the computation of the analysis state more accurate compared to a classical solution.

Section 2 is devoted to explaining the notation. In Section 3, we briefly review the preliminary concept of the 3D-VAR data assimilation scheme. In this setting, an elementary 1D example in the Gaussian domain is presented to elaborate on the underlying assumptions and performance of the methodology in an ideal case. In Section 4, we provide evidence on the sparsity of some important geophysical signals. Exploiting the observed sparsity as an *a priori* knowledge, in Section 5, we cast the variational data assimilation in a Bayesian framework both in the spatial and wavelet domains. The promise of the methodology is demonstrated through an elementary constructed 1D example in that section. Section 6, contains a brief discussion and points out to future research.

2 Notation

We refer to a 2D signal \mathbf{X} as a vector \mathbf{x} , by stacking all the pixels in a fixed order. All vectors are column vectors and $(\cdot)^T$ indicates the transpose. For any vector $\mathbf{x} \in \mathbb{R}^n$, x_i refers to its i^{th} element, where $i \in \{1, \dots, n\}$. The same notation applies to a matrix operator \mathbf{H} and its entries $h_{i,i}$. The standard l_p -norm of \mathbf{x} is denoted by $\|\mathbf{x}\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$, for $p \geq 1$, while the infinity norm is $\|\mathbf{x}\|_\infty = \max_i |x_i|$. For $p < 1$, $\|\mathbf{x}\|_p$ is no longer a norm and hence not convex; nevertheless, we will use the term norm in this case as well, keeping in mind this reservation. By weighted inner product, we denote $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{P}} := \mathbf{x}^T \mathbf{P} \mathbf{y}$ and hence the corresponding weighted Euclidean (quadratic) norm, is $\|\mathbf{x}\|_{\mathbf{P}} = (\mathbf{x}^T \mathbf{P} \mathbf{x})^{1/2}$, where $\mathbf{P} \succ 0$ is a symmetric positive definite matrix. In a linear transformation $\Phi = [\phi_1, \phi_2, \dots, \phi_m] \in \mathbb{R}^{n \times m}$ with $m \geq n$ (e.g., wavelet transform), the columns $\phi_i \in \mathbb{R}^n$ denote the “atoms” whereby any certain class of signals $\mathbf{x} \in \mathbb{R}^n$ can be well approximated by a linear combination of ϕ_i ’s, i.e., $\mathbf{x} \cong \sum_i^m \phi_i c_i = \Phi \mathbf{c}$. The vector \mathbf{x} is said to exhibit a sparse representation in Φ , if the number of (significantly) non-zero elements of the representation coefficients \mathbf{c} , is much smaller than the signal dimension.

3 Variational Data Assimilation

The theory of (recursive) least-squares estimation has been central to the development of the classical data assimilation methodologies. Let the true state of interest at time t be denoted by $\mathbf{x}(t) \in \mathbb{R}^m$, a noisy observation of the state by $\mathbf{y}(t) \in \mathbb{R}^n$, and the background estimate of the state produced by a dynamical model of the underlying physics by $\mathbf{x}_b(t) \in \mathbb{R}^m$. We assume that the model can reproduce an unbiased but noisy estimate of the true state. In other words, it is assumed that the model can resolve the underlying physics in such a way that under a consistent perturbation of the input parameters and states, the ensemble average of the output tends to the true state as the number of ensemble members goes to the infinity, whilst the random deviation of each ensemble member can be well approximated by a Gaussian density. In a more formal setting we have two equations that relate the true state to the background state and observation as follows:

$$\begin{aligned} \mathbf{x}_b &= \mathbf{x} + \mathbf{w} \\ \mathbf{y} &= \mathcal{H}(\mathbf{x}) + \mathbf{v}, \end{aligned} \tag{1}$$

where $\mathbf{w} \sim \mathcal{N}(0, \mathbf{B})$, $\mathbf{v} \sim \mathcal{N}(0, \mathbf{R})$ are uncorrelated Gaussian and the time index is dropped for brevity.

Obviously the goal is now to obtain the so-called analysis state $\mathbf{x}_a \in \mathbb{R}^m$ as the best estimate of the

true state, given the above pair of observation and the background state. From the 3D variational point of view, it amounts to obtaining the analysis state \mathbf{x}_a which minimizes the sum of two quadratic cost functions, each of which quantifies the weighted Euclidean distance of the analysis to the background state \mathbf{x}_b and observation \mathbf{y} :

$$\mathbf{x}_a = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{x}_b - \mathbf{x}\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \|\mathbf{y} - \mathcal{H}(\mathbf{x})\|_{\mathbf{R}^{-1}}^2 \right\}, \quad (2)$$

where the weights are inverse of the error covariance matrices. For now, we assume that the measurement operator $\mathcal{H}(\cdot)$ can be replaced with a linear time invariant $\mathbf{H} \in \mathbb{R}^{n \times m}$ operator. In the context of our study, the incremental formulation for nonlinear measurement operator, see (e.g., [Courtier et al., 1994](#)), which typically arises in direct assimilation of satellite radiance observations, will be briefly discussed in Section 5.

From a statistical estimation point of view, the variational form in equation (2) can be obtained through an ML estimator, $\mathbf{x}_{ML} = \arg \max_{\mathbf{x}} p(\mathbf{y}, \mathbf{x}_b | \mathbf{x})$, where $p(\mathbf{y}, \mathbf{x}_b | \mathbf{x})$ denotes the joint conditional density (Gaussian) of the observation and background state with respect to \mathbf{x} . In this view, the cost function in equation (2) is equivalent to the negative of the log-likelihood function, assuming the background and observation vectors are independent, see (e.g., [Bouttier and Courtier, 2002](#)). Simple algebra and ignoring the constant terms in \mathbf{x} yields a smooth quadratic cost function

$$\begin{aligned} \mathcal{J}(\mathbf{x}) = & \frac{1}{2} \mathbf{x}^T (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \mathbf{x} - \\ & (\mathbf{B}^{-1} \mathbf{x}_b + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y})^T \mathbf{x}, \end{aligned} \quad (3)$$

where the analysis state is its potential unique minimizer, $\mathbf{x}_a = \arg \min_{\mathbf{x}} \mathcal{J}(\mathbf{x})$. Note that the cost function of equation (3) is strictly convex with unique global minimum provided that the Hessian $\nabla^2 \mathcal{J}(\mathbf{x}) = \mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}$ is positive definite which requires that the measurement operator \mathbf{H} be a full rank matrix, see [Olver and Shakiban \(2006, p.160\)](#). This unique minimum can be obtained by setting the first order derivative to zero $\nabla_{\mathbf{x}} \mathcal{J} = 0$

$$\mathbf{x}_a = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} (\mathbf{B}^{-1} \mathbf{x}_b + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y}). \quad (4)$$

Through the Fisher information it can be shown that the obtained analysis in equation (4) can be an *efficient* (unbiased with minimum variance) estimator and its error covariance meets the *Cramer-Rao* lower bound, which is the inverse of the Hessian of $\mathcal{J}(\mathbf{x})$, see (e.g., [Bouttier and Courtier, 2002](#); [Levy, 2008](#), p. 140).

As is evident, the obtained closed form expressions are computationally prohibitive for large scale and ill-conditioned assimilation problems and typically first order iterative approaches (e.g., preconditioned conjugate gradient) are required to efficiently compute the matrix inversions. One of the main advantages of

the variational formalism is its flexibility that for example the analysis state, constrained in a simple feasible and closed polyhedron (e.g., $\mathbf{l} \preceq \mathbf{x} \preceq \mathbf{u}$), can be obtained using the Fast Gradient Projection (FGP) methods developed for large scale quadratic programming problems (e.g., [Nesterov, 1983](#); [Serafini et al., 2005](#)).

Figure 1 shows the result of a 3D-VAR assimilation scheme applied to a stationary first order discrete Markovian process in \mathbb{R}^m , autoregressive (AR-1), where $m = 64$. A true signal (\mathbf{x}) is generated and the background states and observations are obtained by adding Gaussian white noise with the signal-to-noise ratio (SNR) 8 dB and 10 dB respectively, where $\text{SNR} = 20 \log(\sigma_{\mathbf{x}}/\sigma_{\text{noise}})$. Here, to simply resemble the uncaptured subgrid details, we have assimilated a coarse-scale observation signal with half of the size of the original signal. To this end, we first convolved the true signal with an average filter $1/2[+1, +1]$, downsampled the smoothed observations by a factor of 2 and then added the white noise. Notice that in a matrix form, it suffices to define the observation operator \mathbf{H} as a Toeplitz convolution matrix with $h_{i,i} = h_{i+1,i} = 0.5$, $h_{i,j} = 0$ and then decimate the rows by a factor of 2. In other words, for each pair of two grid points in the model space there is only one observation node in the middle, which is assumed to be a noisy measurement of the mean of the true states on those grid points.

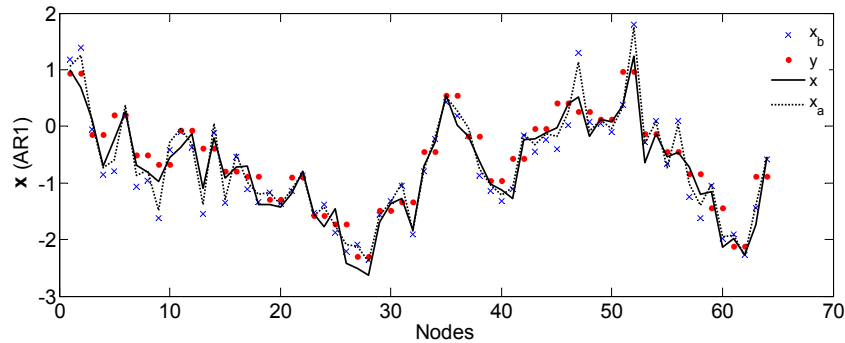


Figure 1: Results of a 3D-Var assimilation in which the true signal (\mathbf{x}) follows a stationary AR(1) discrete Gaussian process. Here we assumed $x_{i+1} = \gamma x_i + \sqrt{1 - \gamma^2} u_i$, where $\gamma = 0.8$ and $u_i \sim \mathcal{N}(0, 1)$. As a result of the assimilation, the analysis exhibits an improved $\text{SNR} = 11.20$ dB.

In the next section, we provide evidence on the non-Gaussian and sparse structure of the fluctuations of some important geophysical signals in terms of their wavelet coefficients. This property is completely ignored in the explained classical formulation of data assimilation and can serve as additional prior knowledge to constrain more and possibly enhance the accuracy of the assimilation results.

4 Sparsity of Geophysical Signals

Many natural signals exhibit a spatial organization of isolated high-intensity areas nested within less active larger-scale regions. This property often translates into a *sparse representation*, that is, a major portion of

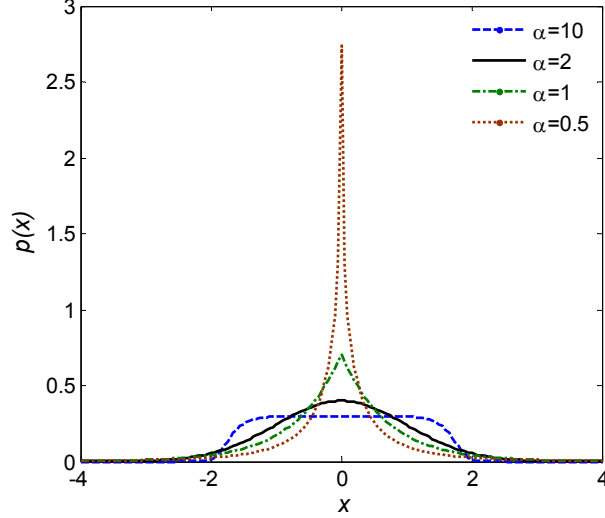


Figure 2: Generalized Gaussian Density with unit standard deviation and various tail parameters.

the signal can be projected onto (near) zero values under an appropriate transformation, while only a few significantly non-zero projection coefficients carry most of the signal energy. Motivated by [Mallat \(1989\)](#), [Huang and Mumford \(1999\)](#) and [Wainwright et al. \(2001\)](#) among many others, it has been recently shown by [Ebtehaj and Fofoula-Georgiou \(2011\)](#) and [Ebtehaj et al. \(2012\)](#) that precipitation reflectivity images exhibit a remarkably sparse representation in a redundant wavelet transform and the distribution of the wavelet coefficients can be well explained by the class of symmetric Generalized Gaussian Distributions (GGD). This family of densities $p(x) \propto \exp(-|x/s|^\alpha)$ with a scale s and tail parameter α spans a wide range of exponentially bounded tail probabilities from a Dirac delta ($\alpha \rightarrow 0$) to a uniform density ($\alpha \rightarrow \infty$) in limiting cases. The Gaussian ($\alpha = 2$) and the Laplace ($\alpha = 1$) densities are also two special cases of this family, see [Figure 2](#).

[Figure 3](#) (upper panels) shows different geophysical signals ranging from very fast evolving dynamical processes such as precipitation and streamflow down to a landscape digital elevation map with a very slow evolving dynamics. Applying a Daubechies wavelet, histograms of the wavelet coefficients share relatively similar thick tail probability distribution, analogous to the GGD, whilst most of the values are (near) zero; see the lower panels of [Figure 3](#). These observations imply that as an a priori knowledge, the wavelet coefficients (generalized fluctuations) of these signals exhibit a sparse representation and can be well explained, at least in part, by the family of GGDs with the tail parameter commonly ranging in $(0, 1]$. Note that, this density is log-concave (i.e., the negative logarithm is a convex function) for $\alpha \geq 1$, and hence the Laplace density ($\alpha = 1$) is the best choice in this family that promotes sparsity while preserving a convex structure. Note that, the concept of sparsity is not restricted only to the wavelet coefficients of physical states with piece-wise smooth structure, such as the examples presented herein. Other (prominent) physical states with smooth

surfaces and trajectories may exhibit sparse representation in other transform domains such as the Fourier or Discrete Cosine Transform (DCT).

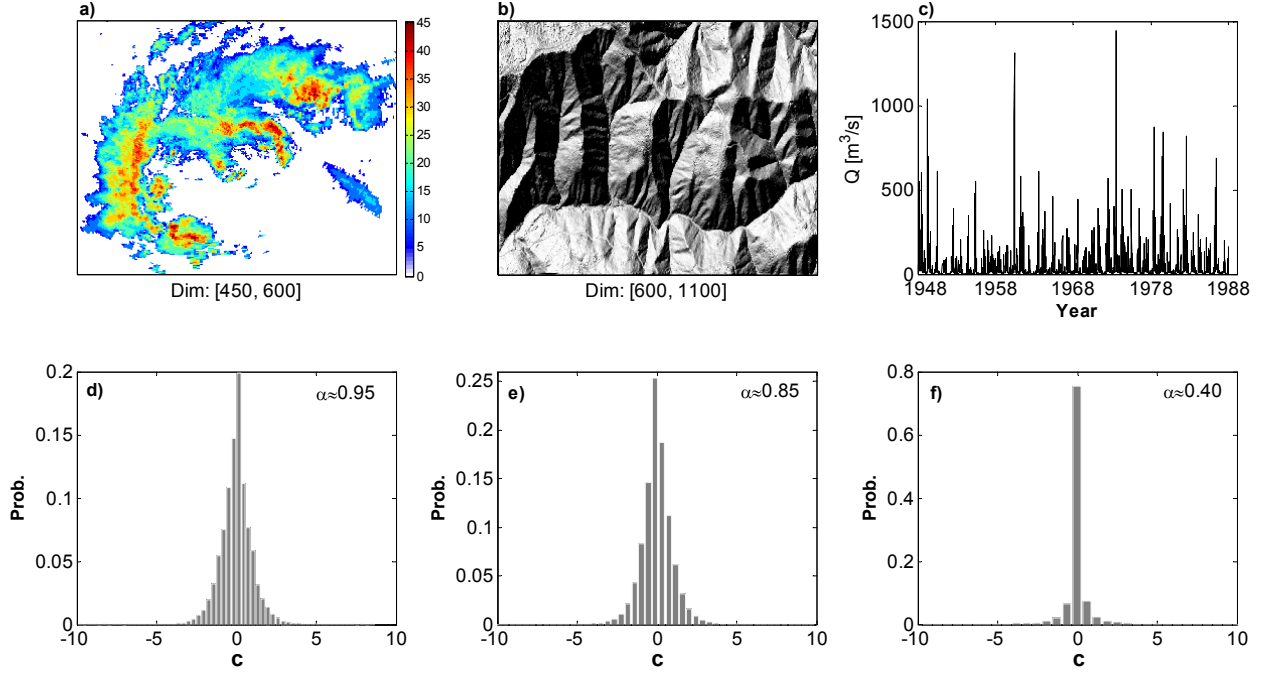


Figure 3: Sparsity of some geophysical signals, top panel from left to right: (a) a level III NEXRAD rainfall reflectivity image in dBZ, over Texas on 1999/03/29 (20:13:00 UTC) at resolution $\sim 1 \times 1$ km; (b) hillshade representation of high resolution lidar topographic data of a small watershed (2.8 km² area) in the Oregon coast range near Coos Bay at resolution $\sim 2 \times 2$ m; and (c) 40 years of daily streamflow signal (1948-1988) of Leaf river basin at Collins station (144 km² draining area), Mississippi. The bottom panels from left to right (d)-to-(f), show the corresponding probability histograms of the standardized wavelet coefficients \mathbf{c} or say the generalized fluctuations of the above images in a probability scale. The fitted tail parameter (α) of the GGD is shown on the top right corner of the lower panel plots.

5 Assimilation with Sparse Priors

Having informative a priori knowledge about the distribution of the analysis state can serve to further constrain the assimilation problem and lead to an improved a posteriori estimate of the analysis from a Bayesian point of view. By definition, the Maximum a posteriori (MAP) estimator of the analysis state is

$$\mathbf{x}_a^+ = \arg \max_{\mathbf{x}} p(\mathbf{x} | \mathbf{x}_b, \mathbf{y}). \quad (5)$$

Applying the Bayes theorem and taking the logarithm, one can obtain

$$\mathbf{x}_a^+ = \arg \min_{\mathbf{x}} \{ \mathcal{J}(\mathbf{x}) - \log p(\mathbf{x}) \}, \quad (6)$$

here $\mathcal{J}(\mathbf{x})$ is the negative log-likelihood function of the classical 3D-VAR, as previously explained. Note that, for a non-informative log-prior or say uniform density for $p(\mathbf{x})$, this expression is exactly equivalent to the cost function of the ML estimator in equation (3).

A general model for the prior distribution, often referred to as the *Gibbs prior*, is given by

$$p(\mathbf{x}) \propto \exp \{-\lambda \mathcal{T}(\mathbf{x})\}, \quad (7)$$

where $\lambda \geq 0$ is a scaling parameter and $\mathcal{T} : \mathbb{R}^m \rightarrow \mathbb{R}$ is a functional mapping from the state space to a real number, see (e.g., [Elad et al., 2007](#)). It follows from equations (6) and (7) that, the MAP estimator of the analysis state is

$$\mathbf{x}_a^+ = \arg \min_{\mathbf{x}} \{\mathcal{J}(\mathbf{x}) + \lambda \mathcal{T}(\mathbf{x})\}, \quad (8)$$

where the non-negative λ acts as a trade-off parameter and plays an important role in the solution of the data assimilation problem. Naturally, small λ weakens the effect of sparse prior and turns the problem into the classical least squares one, while larger λ values, promote a more sparse solution. In this paper, to exploit the underlying sparsity, we suggest two particular choices of the transformation function $\mathcal{T}(\mathbf{x})$ which yields to a sparse-promoting reformulation of the variational data assimilation both in the wavelet and spatial domains.

5.1 Linear Measurement Operator

5.1.1 Wavelet Domain (W3D-VAR)

Let us assume that the analysis state has a projection onto a redundant wavelet transform $\mathbf{x} = \Phi \mathbf{c}$, where the columns of $\Phi \in \mathbb{R}^{m \times k}$ contain the wavelet “atoms”, while $\mathbf{c} \in \mathbb{R}^k$ is the representation coefficients. As is evident, this matrix multiplication is equivalent to an inverse wavelet transform while another matrix $\Psi \in \mathbb{R}^{k \times m}$ represents the forward wavelet transform for obtaining the wavelet coefficients, $\mathbf{c} = \Psi \mathbf{x}$. Given that, the wavelet coefficients of the analysis state exhibit a sparse structure and can be well explained by independent GGD distributions, a relevant choice for the functional mapping $\mathcal{T}(\cdot)$ in the Gibbs prior can take the following form

$$\mathcal{T}(\mathbf{x}) = \sum_i^n (\psi_i x_i)^p = \|\Psi \mathbf{x}\|_p^p = \|\mathbf{c}\|_p^p. \quad (9)$$

Choosing the closest convex representation of $\mathcal{T}(\mathbf{x})$ (i.e., $p = 1$), which is equivalent to assuming a Laplace prior for the wavelet coefficients, it follows that the 3D-VAR can be recast in the wavelet domain as

$$\mathbf{c}_a^+ = \arg \min_{\mathbf{c}} \{\mathcal{J}(\Phi \mathbf{c}) + \lambda \|\mathbf{c}\|_1\}, \quad (10)$$

where \mathbf{c}_a^+ denotes the analysis wavelet coefficients that can be used to reconstruct the analysis state in the physical state space via the inverse wavelet transform, $\mathbf{x}_a^+ = \Phi \mathbf{c}_a^+$. Note that both matrix multiplications, $\Phi \mathbf{c}$ and $\Psi \mathbf{x}$, can be performed very efficiently by the existing fast wavelet transforms, such as the orthogonal wavelet transform (i.e., $\Phi \Psi = \mathbf{I}$) (e.g., [Mallat, 1989](#)). It turns out that, due to its shift invariance property, the class of undecimated wavelet transforms is often preferred in this context, over the traditional discrete orthogonal wavelet transform (e.g., [Coifman and Donoho, 1995](#)). Notice that, assuming $p = 2$ in equation (9) refers to a Gaussian prior for the wavelet coefficients and resembles the so called *Tikhonov regularization* in solving inverse problems. In this case, equation (10) has a closed form solution and this choice of prior is typically very suitable for smooth states.

5.1.2 Spatial Domain (TV3D-VAR)

Following the existence of a sparse structure in the wavelet coefficients or say generalized fluctuations of a geophysical signal \mathbf{x} , another choice for the functional mapping $\mathcal{T}(\mathbf{x})$ in the Gibbs prior, is the Total Variation (TV) semi-norm of \mathbf{x} , which leads to obtaining the analysis as

$$\mathbf{x}_a^+ = \arg \min_{\mathbf{x}} \{ \mathcal{J}(\mathbf{x}) + \lambda \|\mathbf{x}\|_{\text{TV}} \}. \quad (11)$$

Two popular choices for the discrete TV semi-norm (e.g., [Rudin et al., 1992](#); [Beck and Teboulle, 2009b](#)) are: the isotropic one

$$\|\mathbf{x}\|_{\text{TV}_I} = \sum_{i=1}^n \sqrt{(\nabla_h x_i)^2 + (\nabla_v x_i)^2}, \quad (12)$$

and the l_1 -based

$$\|\mathbf{x}\|_{\text{TV}_{l_1}} = \sum_{i=1}^n (|\nabla_h x_i| + |\nabla_v x_i|), \quad (13)$$

where, $\nabla_h x_i$ and $\nabla_v x_i$ are horizontal and vertical first order differences at pixel i , respectively. Note that obtaining the optimal solution of the TV3D-VAR cost function is more involved than the W3D-VAR as the TV semi-norm is not a separable functional.

5.2 Nonlinear Measurement Operator

By first order linearization of the measurement operator in equation (2) and a change of variable $\delta \mathbf{x} = \mathbf{x} - \mathbf{x}_b$, the classical 3D-Var in an incremental form is typically reformulated as ([Courtier et al., 1994](#)),

$$\mathcal{J}(\delta \mathbf{x}) = \frac{1}{2} \|\delta \mathbf{x}\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \|\delta \mathbf{y} - \mathbf{H} \delta \mathbf{x}\|_{\mathbf{R}^{-1}}^2, \quad (14)$$

where $\delta \mathbf{y} = \mathbf{y} - \mathcal{H}(\mathbf{x}_b)$, and here \mathbf{H} is a suitable linear approximation (e.g., Jacobian) of $\mathcal{H}(\mathbf{x})$ in a small neighborhood around \mathbf{x}_b . Finding $\delta \mathbf{x}_a$ as the minimizer of equation (14), the analysis can be obtained by, $\mathbf{x}_a = \mathbf{x}_b + \delta \mathbf{x}_a$. Having the sparse prior assumption on the transformed increments of the analysis state, it naturally leads to a possible choice for an incremental MAP estimator as follows:

$$\delta \mathbf{x}_a^+ = \arg \min_{\mathbf{x}} \{ \mathcal{J}(\delta \mathbf{x}) + \lambda \|\Psi \delta \mathbf{x}\|_1 \}, \quad (15)$$

where, here Ψ is referred to an invertible transformation (e.g., wavelet) that sparsifies the increments. We again emphasize the fact that the above proposed formulation seems intuitively suitable for states with sparse increments under a proper transformation. Of course, complementary case studies and further empirical evidence are required for a thorough conclusion about the proper selection of the transformation.

Notice that, although the proposed data assimilation formalism in equations (10), (11) and (15) is convex, the prior terms are not differentiable and hence the cost function is *non-smooth*. In this case classical (first order) gradient based methods are no longer applicable. Several optimization techniques have been recently proposed to deal with large-scale non-smooth convex cost functions similar to that in equation (8), where \mathcal{J} is smooth and \mathcal{T} is a non-smooth convex function. A large effort has been devoted on using efficient interior point algorithms for this particular type of cost functions in large scale problems (e.g., [Goldfarb and Yin, 2005](#); [Kim et al., 2007](#)). However, very recently, accelerated proximal gradient methods have received significant attention due to their fast convergence rate and simplicity (e.g., [Nesterov, 2007](#); [Figueiredo et al., 2007](#); [Bioucas-Dias and Figueiredo, 2007](#); [Beck and Teboulle, 2009a,b](#)).

5.3 Non-Gaussian Error

All of the presented formulations so far have been focused on the fact that the model $\mathbf{v} \in \mathbb{R}^m$ and measurement $\mathbf{w} \in \mathbb{R}^n$ error terms can be well explained by a multivariate Gaussian distribution as the most dominant error probability model. Sometimes the distribution of the error is symmetric with tails markedly thicker than the Gaussian case, analogous to the GGD family. In this case, it can be shown that naturally the ML estimator in equation (2) is,

$$\underset{\mathbf{x}}{\text{minimize}} \left\| \mathbf{B}^{-1/2} (\mathbf{x}_b - \mathbf{x}) \right\|_{p_1}^{p_1} + \left\| \mathbf{R}^{-1/2} (\mathbf{y} - \mathcal{H}(\mathbf{x})) \right\|_{p_2}^{p_2}. \quad (16)$$

Notice that for $p_i = 2$, $i \in \{1, 2\}$, the problem in equation (16) is equivalent to the classical 3D-VAR cost function and is convex for all $p_i \geq 1$. As explained before, it can be concluded that the Laplace model for the error is the thickest tail probability that can be considered while preserving convexity of the cost function.

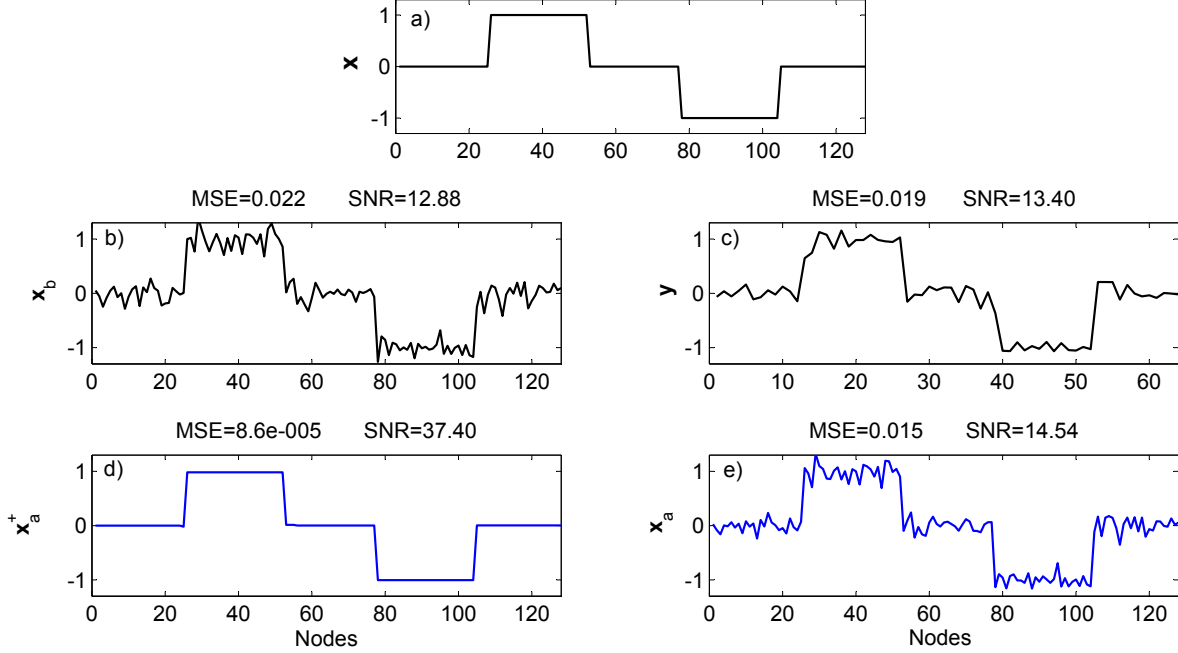


Figure 4: Demonstration of the promise of the 3D-VAR with sparse prior for a 1D example. From top-to-bottom, (a) the true signal \mathbf{x} , (b) the noisy background \mathbf{x}_b , (c) the low-resolution and noisy observation \mathbf{y} , (d) the obtained analysis with prior \mathbf{x}_a^+ , and (e) the analysis \mathbf{x}_a that resulted from the classical 3D-VAR.

Obviously, the above ML estimator in a Generalized Gaussian noise environment can be further extended to the MAP estimator by adding the prior term, as previously explained.

6 A 1D Synthetic Example

In this section, by no means we intend to solve a real data assimilation problem but only to demonstrate the promise of the proposed formulations and specifically the role of the prior on the “analysis phase”. To this end, we focused on a very simple piece-wise constant 1D example with an extremely sparse structure in its first order differences, see Figure 4. As is evident, the first order differences of this signal exhibit a marked sparsity with only four non-zero elements at the jump discontinuities. Notice that, in this case the wavelet l_1 -based and the explained TV-based priors become analogous provided that the wavelet dictionaries contain the Haar “atoms”. To simplify and be more instructive, we have chosen a first order differencing operator for obtaining a sparse representation.

In this case, the 3D-VAR with sparse prior can be recast in the following simple form

$$\mathbf{x}_a^+ = \arg \min_{\mathbf{x}} \{ \mathcal{J}(\mathbf{x}) + \lambda \|\mathbf{D}\mathbf{x}\|_1 \}, \quad (17)$$

where, $\mathbf{D} \in \mathbb{R}^{m \times m}$ is the first order differencing operator with $d_{i,i} = 1$, $d_{i,i-1} = -1$ and $d_{i,j} = 0$. To obtain

the analysis in equation (17) here we follow a quadratic reformulation of the problem, as studied by [Chen et al. \(1998\)](#) and [Figueiredo et al. \(2007\)](#) and references therein. To this end, by a change of variable $\mathbf{z} = \mathbf{D}\mathbf{x}$ one can obtain

$$\mathbf{z}_a = \arg \min_{\mathbf{x}} \{ \mathcal{J}(\mathbf{D}^{-1}\mathbf{z}) + \lambda \|\mathbf{z}\|_1 \}, \quad (18)$$

where $\mathbf{z} \in \mathbb{R}^m$ can be split as $\mathbf{z} = \mathbf{u} - \mathbf{v}$, with $u_i = \max(z_i, 0)$ and $v_i = \max(-z_i, 0)$. Accordingly, the l_1 -prior term can be written in a linear form as $\|\mathbf{z}\|_1 = \mathbf{1}_m^T \mathbf{u} + \mathbf{1}_m^T \mathbf{v}$, where $\mathbf{1}_m = [1, 1, \dots, 1]^T \in \mathbb{R}^m$. Augmenting \mathbf{u} and \mathbf{v} in $\mathbf{w} = [\mathbf{u} \ \mathbf{v}]^T$, equation (18) can be recast in the following constrained quadratic programming (QP)

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \frac{1}{2} \mathbf{w}^T \begin{bmatrix} \mathbf{C} & -\mathbf{C} \\ -\mathbf{C} & \mathbf{C} \end{bmatrix} \mathbf{w} + \left(\lambda \mathbf{1}_{2m} + \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix} \right)^T \mathbf{w} \\ \text{subject to} \quad & \mathbf{w} \succcurlyeq 0, \end{aligned} \quad (19)$$

where, $\lambda \geq 0$, $\mathbf{C} = \mathbf{D}^{-T} (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \mathbf{D}^{-1}$ and $\mathbf{b} = -\mathbf{D}^{-T} (\mathbf{B}^{-1} \mathbf{x}_b + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y})$. Through sub-differential analysis of equation (18), it can be shown that for $\lambda \geq \|\mathbf{b}\|_\infty$ the unique minimum of equation (19) is a zero vector with maximum possible sparsity. Here, we adopt $\lambda = 0.1 \|\mathbf{b}\|_\infty$ as suggested by [Kim et al. \(2007\)](#).

The example signal $\mathbf{x} \in \mathbb{R}^{128}$ ($\sigma_{\mathbf{x}} \cong 0.65$) is a composition of two rectangular step functions. The background signal $\mathbf{x}_b \in \mathbb{R}^{128}$ in Figure (4b) is generated via adding a white Gaussian noise ($\sigma_w \cong 0.15$, SNR $\cong 12.88$ dB). Here, we assimilated a low-resolution and noisy version of the true signal as the observation $\mathbf{y} \in \mathbb{R}^{64}$ into the background signal. To this end, the observation operator $\mathbf{H} \in \mathbb{R}^{64 \times 128}$ is properly designed as explained in Section 2 and then a Gaussian white noise ($\sigma_{\mathbf{y}} = 0.10$, SNR $\cong 13.40$ dB) is added to the downgraded version of the true signal, see Figure (4c). In this example, we used the Gradient Projection with backtracking line search (i.e., Armijo Rule) for solving the constrained QP in equation (19), see ([Bertsekas, 1999](#), p.230). Furthermore, after obtaining \mathbf{z}_a , we optionally recalculated the magnitude of its nonzero elements by solely minimizing $\mathcal{J}(\mathbf{D}^{-1}\mathbf{z})$, the least squares part of equation (18), constrained to the support set of \mathbf{z}_a , $\mathcal{S} = \text{supp}(\mathbf{z}_a)$. In other words, we assumed that the zero elements of \mathbf{z}_a are fixed and then calculated the magnitude of its non-zero elements, that is $\mathbf{C}_{\mathcal{S}}^\dagger \mathbf{b}$, where $\mathbf{C}_{\mathcal{S}}$ is a sub-matrix that contains those columns of \mathbf{C} associated to the support set \mathcal{S} and $(\cdot)^\dagger$ is the Pseudo-inverse.

The results in Figure 4 show the remarkable role of the sparse promoting prior on the quality of the estimated analysis state. It is clear that the estimation quality metrics have been slightly improved by solving a classical 3D-VAR assimilation problem; however, it led to over-fitted estimation. The result of the

new proposed formulation is close to an exact solution in this simple example and seems very promising by outperforming the classical 3D-VAR with more than two orders of magnitude in the SNR, which is a logarithmic metric. These results demonstrate that the error in the observation and background signal has been well suppressed while the discontinuities of the signal are also well recovered due to the incorporation of the prior.

7 Discussion and Conclusion

We introduced a new formalism for the variational data assimilation which takes into account a priori knowledge about the underlying statistical structure of the state in a transformed domain and showed the preliminary promise of the proposed methodology through a synthetic 1D example. Although the formulation is presented for a 3D-VAR setting, it can be extended to a 4D-VAR context. In general, we can argue that the proper selection of the prior term typically yields a better error (noise) suppression, while the underlying structure of the state (e.g., discontinuities) can also be preserved. Although the focus of this study has been on sparsity of the state fluctuations and wavelet coefficients, the role of the prior can be extended to other transformed domains such as the Fourier or DCT which are intuitively more suitable for smooth physical states.

Study of the efficient proximal gradient methods for full scale and ill-conditioned data assimilation problems with non-smooth prior can be of particular interest for future research in exploring the advantages of the proposed formulations in environmental predictability. The proposed W3D-VAR and TV3D-VAR are nonlinear estimators and hence, estimation of the analysis error covariance is a challenge and needs to be thoroughly investigated. As closed form expressions are not readily available for the covariance of these nonlinear estimators, randomization (e.g., bootstrapping) via ensemble techniques seems a viable approach for further study in this respect.

Acknowledgments

This work has been mainly supported by NASA-GPM award NNX07AD33G, and an Interdisciplinary Doctoral Fellowship (IDF) of the University of Minnesota Graduate School. The second author also wishes to acknowledge the support provided by the Ling Chaired Professorship.

References

- Beck, A., and M. Teboulle (2009a), A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.*, *2*(1), 183–202, doi:10.1137/080716542.
- Beck, A., and M. Teboulle (2009b), Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, *IEEE Trans. Image. Process.*, *18*(11), 2419–2434, doi:10.1109/TIP.2009.2028250.
- Bertsekas, D. P. (1999), *Nonlinear Programming*, 2nd ed., 794 pp., Athena Scientific, Belmont, MA.
- Bioucas-Dias, J., and M. Figueiredo (2007), A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration, *IEEE Trans. Image. Process.*, *16*(12), 2992–3004, doi:10.1109/TIP.2007.909319.
- Bouttier, F., and P. Courtier (2002), Data assimilation concepts and methods, *Meteorological training course lecture series. ECMWF*, p. 59.
- Chen, S. S., D. L. Donoho, and M. A. Saunders (1998), Atomic decomposition by basis pursuit, *SIAM J. Sci. Comput.*, *20*, 33–61.
- Coifman, R., and D. Donoho (1995), Translation invariant de-noising, *Lecture Notes in Statist.*, *103*, 125–150.
- Courtier, P., J.-N. Thépaut, and A. Hollingsworth (1994), A strategy for operational implementation of 4d-var, using an incremental approach, *Quart. J. Roy. Meteor. Soc.*, *120*(519), 1367–1387, doi:10.1002/qj.49712051912.
- Daley, R. (1993), *Atmospheric data analysis*, 472 pp., Cambridge University Press.
- Ebtehaj, A. M., and E. Foufoula-Georgiou (2011), Statistics of precipitation reflectivity images and cascade of gaussian-scale mixtures in the wavelet domain: A formalism for reproducing extremes and coherent multiscale structures, *J. Geophys. Res.*, *116*, D14110, doi:10.1029/2010JD015177.
- Ebtehaj, A. M., E. Foufoula-Georgiou, and G. Lerman (2012), Sparse regularization for precipitation down-scaling, *J. Geophys. Res.*, *117*, D08107 doi:10.1029/2011JD017057, .
- Elad, M., P. Milanfar, and R. Rubinstein (2007), Analysis versus synthesis in signal priors, *Inverse Problems*, *23*(3), 947.
- Evensen, G. (2007), *Data Assimilation: The Ensemble Kalman Filter*, 307 pp., Springer.

- Figueiredo, M., R. Nowak, and S. Wright (2007), Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, *IEEE J. Sel. Topics Signal Process.*, 1(4), 586–597, doi:10.1109/JSTSP.2007.910281.
- Ghil, M., and P. Malanotte-Rizzoli (1991), Data assimilation in meteorology and oceanography, pp. 141 – 266, Elsevier, doi:10.1016/S0065-2687(08)60442-2.
- Goldfarb, D., and W. Yin (2005), Second-order cone programming methods for total variation-based image restoration, *SIAM J. Sci. Comput.*, 27(2), 622.
- Huang, J., and D. Mumford (1999), Statistics of natural images and models, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 637–663, doi:10.1109/CVPR.1999.786990.
- Kalnay, E. (2003), *Atmospheric modeling, data assimilation, and predictability*, 341 pp., Cambridge University Press, New York.
- Kim, S.-J., K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky (2007), An interior-point method for large-scale l_1 -regularized least squares, *IEEE J. Sel. Topics Signal Process.*, 1(4), 606–617, doi:10.1109/JSTSP.2007.910971.
- Levy, B. C. (2008), *Principles of Signal Detection and Parameter Estimation*, 1 ed., 639 pp., Springer Publishing Company, Incorporated, New York, USA, doi:10.1007/978-0-387-76544-0.
- Mallat, S. (1989), A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7), 674–693, doi:10.1109/34.192463.
- Nesterov, Y. E. (1983), A method for solving the convex programming problem with convergence rate $O(1/k^2)$, *Dokl. Akad. Nauk SSSR*, 269, 543–547.
- Nesterov, Y. E. (2007), Gradient methods for minimizing composite objective function, *Tech. rep.*, CORE.
- Olver, P., and C. Shakiban (2006), *Applied linear algebra*, 714 pp., Prentice Hall, Upper Saddle River, NJ 07458.
- Rudin, L. I., S. Osher, and E. Fatemi (1992), Nonlinear total variation based noise removal algorithms, *Phys. D*, 60(1), 259–268, doi:10.1016/0167-2789(92)90242-F.
- Serafini, T., G. Zanghirati, and L. Zanni (2005), Gradient projection methods for quadratic programs and applications in training support vector machines, *Optim. Methods Softw.*, 20(2-3), 353–378, doi:10.1080/10556780512331318182.

- Talagrand, O., and P. Courtier (1987), Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory, *Quart. J. Roy. Meteor. Soc.*, *113*(478), 1311–1328.
- Wainwright, M. J., E. P. Simoncelli, and A. S. Willsky (2001), Random cascades on wavelet trees and their use in analyzing and modeling natural images, *Appl. Comput. Harmon. Anal.*, *11*(1), 89 – 123, doi:10.1006/acha.2000.0350.
- Zhou, Y., D. McLaughlin, and D. Entekhabi (2006), Assessing the performance of the ensemble kalman filter for land surface data assimilation, *Mon. Weather Rev.*, *134*(8), 2128–2142.